TECHNICAL ASPECTS OF COMPUTER PROCESSING OF NATURAL LANGUAGE INFORMATION

Ihor Kulchytskyy
Institute of Computer Science and Information Technology, Lviv Polytechnic National University,
S. Bandery Str., 12, Lviv, 79013, UKRAINE

The article deals with technical problems of natural language information processing by computer.

A large part of the modern world society — is its information space based upon the physical media, the lion's share of which exists in electronic verbal form. This space from one (let's call it “humanitarian”) perspective is explored by the specialists in cultural studies, sociology, philosophy, journalism, philology and library science etc., and from the other (“technical”) perspective — by the specialists in social communication (in our opinion, the technical component prevails in their studies) and informatics. The sphere of interests of the latter ones — is mostly internal representation of information in the electronic environment, its protection and transmission via information channels, technical side of information retrieval, and creation of the necessary linguistic software. However, experts of these two groups rarely contact with each other on a scientific level. As a result — each group has superficial and sometimes false representation of another’s group problem, which, in turn, does not contribute at all to the functioning efficiency of the information space. The purpose of this article — is to partially eliminate this barrier.

In this article the specifics of natural language texts processing by computer are determined, the basic methods of their internal representation are overviewed, and the problems that accompany the creation and processing of texts in electronic form are presented. As a result, the following conclusions were made.

During automated processing of electronic texts computer deals not with symbols but with their bit sequences, organized in special codes.

There are a number of established and partially standardized code tables. Therefore, a set of texts, obtained from various sources, which is planned either to be used in scientific researches, or to be put in a particular repository, must be normalized — reduce all texts to one code table, check texts against correct punctuation (for example, check whether an apostrophe is always indicated as the same character), remove extra characters (such as multiple spaces in a row, empty paragraphs, etc.), unify the means and ways of text formatting and so on.

When typing texts it is necessary to use the polygraph rules of typing texts. To this end, the appropriate professionals should generalize these rules and harmonize them with the available code tables (ideally Unicode) and features of the most popular text editing programs and bring them to the widest possible audience.

Keywords – information space, information environment, information society, computer technologies, character encoding, encoding standards.